Domains and functions

Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

24 | 2019
Varia

Domains and Functions: A Two-Dimensional Account of Discourse Markers

Ludivine Crible and Liesbeth Degand

Electronic version
URL: http://journals.openedition.org/discours/9997
DOI: 10.4000/discours.9997
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Ludivine Crible and Liesbeth Degand, « Domains and Functions: A Two-Dimensional Account of Discourse Markers », *Discours* [Online], 24 | 2019, Online since 30 October 2019, connection on 01 November 2019. URL : http://journals.openedition.org/discours/9997 ; DOI : 10.4000/discours.9997

discours
Revue de linguistique, psycholinguistique et informatique
http://journals.openedition.org/discours/

Domains and Functions: A Two-Dimensional Account of Discourse Markers

Ludivine Crible
Université catholique de Louvain – Fonds de la recherche scientifique (FRS-FNRS)

Liesbeth Degand
Université catholique de Louvain

Ludivine Crible, Liesbeth Degand, « Domains and Functions: A Two-Dimensional Account of Discourse Markers », *Discours* [En ligne], 24 | 2019, mis en ligne le 30 octobre 2019.

URL : http://journals.openedition.org/discours/9997

Titre du numéro : *Varia*
Coordination : Anne Le Draoulec & Josette Rebeyrolle

Date de réception de l'article : 05/12/2018
Date d'acceptation de l'article : 07/05/2019

Domains and Functions: A Two-Dimensional Account of Discourse Markers

Ludivine Crible
Université catholique de Louvain – Fonds de la recherche scientifique (FRS-FNRS)

Liesbeth Degand
Université catholique de Louvain

Discourse markers and their functions have been modeled through a large number of very diverse frameworks. Most of these models target written language and the discourse relations that hold between sentences. In this paper, we present, assess and apply a new annotation taxonomy that targets discourse markers (instead of discourse relations) in spoken language and addresses their polyfunctionality in an alternative way. In particular, its main innovative feature is to distinguish between two independent layers of semantic-pragmatic information (i.e., domains and functions) which, once combined, provide a fine-grained disambiguation of discourse markers. We compare the affordances of this model to existing proposals, and illustrate them with a corpus study. A sample of conversational French containing 423 discourse marker tokens was fully analyzed by two independent annotators. We report on inter-annotator agreement scores, as well as quantitative analyses of the distribution of domains and functions in the sample. Both powerful and economical, this proposal advocates a flexible and modular approach to discourse analysis, and paves the way for further corpus-based studies on the challenging category of discourse markers.

Keywords: discourse markers, corpus annotation, speech, polyfunctionality, domains, French

Les marqueurs du discours et leurs fonctions ont fait l'objet de modélisations nombreuses et variées. La plupart de ces modèles portent sur l'écrit et sur les relations discursives entre énoncés. Dans cet article, nous présentons, évaluons et appliquons un nouveau modèle d'annotation qui porte sur les marqueurs du discours (et non sur les relations discursives) à l'oral, offrant une perspective nouvelle sur la polyfonctionnalité des marqueurs. Sa caractéristique la plus innovante est de définir deux couches indépendantes d'information sémantico-pragmatique (à savoir, domaines et fonctions) qui, une fois combinées, fournissent une désambiguïsation fine des marqueurs du discours. Nous comparons les apports de ce modèle à d'autres approches existantes et les illustrons dans une étude de corpus. Un échantillon de français conversationnel contenant 423 marqueurs du discours a été entièrement analysé par deux annotateurs. Nous analysons les scores d'accord inter-annotateurs, ainsi que la distribution des domaines et des fonctions dans l'échantillon. À la fois puissant et économique, ce modèle prône une approche flexible et modulaire de l'analyse du discours, et jette les bases pour de futures études de corpus sur la catégorie complexe des marqueurs du discours.

Mots clés : *marqueurs du discours, annotation de corpus, oral, polyfonctionnalité, domaines, français*

We would like to thank the anonymous reviewers and the editors for their careful and insightful suggestions. Any remaining errors are ours.

1. Introduction

1 In human communication, discourse is where the magic happens. It is through markers of structure and interaction that speakers convey not only the coherence of

their intended message but also their attitude towards this message and towards the interlocutor. Such expressions are called "discourse markers" (henceforth DMs) and have been extensively studied in the past thirty years through a range of theoretical and methodological paradigms, starting from Schiffrin's (1987) seminal study. She defines DMs as "sequentially dependent elements which bracket units of talk" (Schiffrin, 1987: 31), a definition which encompasses both "connectives" (e.g., *and*, *but*, *because*, *actually*) and pragmatic particles more specific to speech (e.g., *well*, *I mean*, *you know*). However, the functions of DMs go much further than this "bracketing" role, as Schiffrin herself acknowledges with her five "planes of talk", i.e., dimensions of the interaction that are targeted by various (functions of) DMs. Thus, the DM can refer to the ideational structure (linking propositions), the action structure (linking speech acts), the exchange structure (taking or yielding turns), the information state (organizing knowledge) or the participation framework (establishing speaker relations).

2 Schiffrin's (1987) model, while influential and widespread (e.g., Buysse, 2012; Sprott, 1992), is however not specifically designed for systematic corpus application and remains qualitative in nature. Alternative approaches have been proposed that vary in the number and types of values that are distinguished in the model, as well as in the method (automatic vs. manual) and data type for which they are intended (spoken vs. written corpora). Among them, the Penn Discourse Treebank 2.0 (Prasad et al., 2008) and Rhetorical Structure Theory (Mann & Thompson, 1988) are particularly well developed, as they have been applied to speech and writing in multiple languages. Other approaches to the annotation and description of DMs and discourse relations refer instead to discourse segmentation (Briz & Pons, 2010) or to basic cognitive primitives (Sanders et al., 1992) in order to tackle DMs' challenging functional variation.

3 In this paper, we present, assess and apply a new annotation model for the functions of DMs in spoken languages. It is an extensive revision of Crible's (2017) taxonomy based on methodological suggestions in Crible and Degand's (2019) annotation experiments. Like these previous proposals, the present model targets the whole DM category (as opposed to fine-grained case studies), covers functions that apply to both speech and writing, and aims at high reliability, even though annotation remains a challenging and somewhat subjective task (Spooren & Degand, 2010). It also shares with Redeker (1990), González (2005), Maschler (2009) or Cuenca (2013) the assumption that discourse functions can be grouped in three or four "domains", i.e., macro-functions which roughly correspond to the speaker's intention and degree of involvement. However, it stands out from its predecessors by offering a two-dimensional account of DM polyfunctionality, whereby functions and domains are independent, thus vouching for an economical yet powerful model for systematic discourse analysis.

4 While the basic principles of our model were already sketched in Crible and Degand (2019), where we discussed the impact of annotators' expertise on reliability, in this paper we present the final taxonomy after further stages of testing and

revisions. The objectives of this paper are therefore the following: firstly, to serve as a reference paper providing operational annotation guidelines for all the values in the model; secondly, to compare this model with other proposals, thus highlighting their different benefits and complementarity; finally, to illustrate its affordances on a sample of spoken French corpus data, and to discuss original findings on the distribution and combination of DM functions and domains in this language.

5 In the next section, we will carefully review a selection of previous proposals to discourse annotation. We will then introduce our new model, operationally defining the four domains and fifteen functions and how they combine. In Section 4, we present the corpus data to which this model was applied and the inter-annotator agreement measures that we reached on this sample. We also report quantitative findings on the distribution of domains and functions in conversational French, which will illustrate the affordances of this model. Finally, we conclude by discussing the inter-relation between an annotation model and its research objectives.

2. Previous approaches to discourse annotation

6 In this section, we focus on three types of approaches to discourse functions and annotation, which were selected because of the influence they had on the model we are introducing in this paper. They are also representative of quite distinct traditions in the field, as they adopt different perspectives to polyfunctionality and follow different research agendas.

2.1. Hierarchical inventories

7 One of the most influential and widespread models of annotation for discourse relations is the Penn Discourse Treebank (henceforth PDTB) in its various versions, the latest being 3.0 (Prasad et al., 2018). In the PDTB, discourse relations such as Reason or Concession have been manually annotated, regardless of whether an explicit DM was used to signal the relation. The hierarchical taxonomy includes four semantic classes (TEMPORAL, CONTINGENCY, COMPARISON, EXPANSION), which are further distinguished in one or two levels comprising more specific values (e.g., TEMPORAL includes Asynchronous which includes Precedence and Succession). The sense hierarchy is represented in Figure 1. We can see that some functions do not have a level-3 value (e.g., Synchronous, Contrast), while others are also distinguished at level 2 (e.g., Condition vs. Negative condition). Level-3 distinctions only apply to asymmetric relations and "capture the directionality of the arguments" (Prasad et al., 2018: 90), such as Arg1-as-cond vs. Arg2-as-cond[1].

1. There are two exceptions to this principle: Negative-Result and Arg2-as-Negative-Goal relations do not have an asymmetric variant ("Negative-Cause" or "Arg1-as-Negative-Goal" do not appear on the taxonomy). In addition, the latter is no longer included in the final version of PDTB 3.0 annotation manual (publicly available at: https://catalog.ldc.upenn.edu/docs/LDC2019T05/).

Temporal	Synchronous	—
	Asynchronous	Precedence
		Succession

Contingency	Cause +/-β, +/-ζ	Reason
		Result
		Negative-result*
	Condition +/-ζ	Arg1-as-cond
		Arg2-as-cond
	Negative condition +/-ζ	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal

Comparison	Contrast	—
	Similarity	—
	Concession +/-ζ	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	—
	Disjunction	—
	Equivalence	—
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Figure 1 – Hierarchy of senses in the PDTB 3.0 (Prasad et al., 2018: 90)

8 The main principle of the PDTB is that the annotators can stop at a higher level if they cannot decide on a more specific value because of an ambiguity or a disagreement between annotators. Prasad et al. (2008) reported on an inter-annotator agreement score of 84% for level-2 senses using the PDTB 2.0. In this previous version, some DMs (or "explicit connectives" in their terminology) were assigned up to thirty different labels (this is the case for *but* or *when*).

9 The PDTB has been applied to written data in many different languages and, to a much smaller extent, to spoken data as well (e.g., Tonelli et al., 2010, in Italian; or Demirşahin & Zeyrek, 2014, in Turkish). It targets discourse *relations* rather than discourse *markers*: even though its approach is relatively more connective-oriented than other frameworks (for instance, a connective is always reconstructed even in implicit relations), the PDTB aims at a comprehensive coverage of discourse relations, regardless of their marking (explicit, implicit, or marked by alternative lexicalizations). As a result, it does not focus on the multiple functions that some DMs may perform beyond discourse relations (such as topic-shifting, turn-taking, repairing, etc.), but rather provides a comprehensive description of discourse relations and their various forms of marking. It is referential in the field as one of the most widespread frameworks (for recent studies, see e.g., Lee et al., 2016; Zhang et al., 2016).

2.2. Cognitive primitives

10 Another important framework is the Cognitive approach to Coherence Relations (henceforth CCR) based on the seminal proposal by Sanders et al. (1992 and 1993). The particularity of the taxonomy is to decompose discourse relations into four binary features: basic operation (additive vs. causal), order of the segments (basic vs. non-basic), polarity (positive vs. negative) and source of coherence (objective vs. subjective). This approach targets psychological plausibility rather than complete descriptive adequacy (Sanders et al., 1992: 4), even though the assumption is that most relations can be described along these cognitive primitives. As opposed to most other frameworks, in the CCR, relations and their markers are not assigned end-labels such as "contrast" or "consequence" but a combination of four primitives (e.g., causal, positive, basic, objective). In recent work, a proposal has been made to add "missing" primitives such as temporality in order to reach better linguistic and cognitive coverage (Evers-Vermeul et al., 2017).

11 This approach aims at maximal replicability, as making binary decisions (e.g., additive vs. causal) is considered more reliable and more robust than choosing from long lists of labels (for an annotation experiment, see Scholman et al., 2016). More importantly, each primitive has been associated with psycholinguistic evidence from experimental and corpus-based studies, which showed that these binary distinctions are indeed cognitively relevant and used by language users in processing or in acquisition. For instance, positive relations are acquired before negative ones, and additives before causals (Evers-Vermeul & Sanders, 2009).

12 Similarly to the PDTB, the CCR aims at accounting for discourse/coherence *relations* rather than their explicit markers. It follows that, by design, the CCR does

not cover additional functions of DMs besides discourse relations: for instance, topic relations are considered as orthogonal to discourse relations, and are therefore not included in the taxonomy (Sanders et al., 2018: 63). The authors acknowledge that some distinctions are lost with this binary system, as several relations may share exactly the same features: for instance "Conjunction" and "Instantiation" from the PDTB 2.0 would both be additive, positive and objective (order of segments does not apply). Finally, the dichotomy between objective and subjective relations has been extended to all relations, in contrast to the PDTB 3.0, where it only applies to Cause, Condition, Negative-condition and Concession relations (marked as "belief" and/or "speech-act"). Still, this restrains the variation of discourse relations to these two options, whereas DM functions are often seen as having three or even four variants in other frameworks (see Section 3.1 below).

2.3. Discourse segmentation

13 A third approach that we will briefly mention starts from the angle of segmentation into various units and assumes that the functions of DMs stem from the type of unit in which they occur and their position in this unit. This approach is represented for instance by the Val.Es.Co research group (e.g., Briz & Pons, 2010) who work on conversational Spanish. They distinguish between eight unit types: discourse, dialogue, exchange, turn alternation, intervention, turn, act, subact. The first four are dialogical, the latter monological. The aim of this segmentation approach is to provide an exhaustive, recursive and hierarchical account of spoken discourse structure (Val.Es.Co Group, 2014).

14 DMs themselves are defined as either turns (e.g., some interjections), adjacent subacts (e.g., *well*) or parts of subacts (all conjunctions). They can be initial, medial or final with respect to the other units and their function is either textual, interpersonal or modal. Unit and position constrain the function of the DM, so that a DM in final position of an intervention will likely be interpersonal.

15 Contrary to the previous two frameworks reviewed so far, this segmentation approach is designed specifically for spoken DMs, and aims at capturing the relationship between discourse structure and DM functions. The three-fold functional distinction is adequate to meet this goal, even though it might not be fine-grained enough for other research questions. This objective also explains why the Val.Es.Co model assumes a strongly deterministic relationship between the syntagmatic position of a DM and its function, when other studies have shown that the variation of DMs is not always predictable and systematic (see e.g., Degand, 2014; Heim, 2019, on the limits of the association between peripheral use and subjective or intersubjective meanings).

3. The proposal: a two-dimensional account of DMs

16 Our selective literature review has uncovered the need for an annotation model that accounts for the specific characteristics and polyfunctionality of DMs in speech, with

a view to providing comprehensive quantitative studies of this challenging category across various (spoken) languages. We now turn to the presentation of our proposal, starting from a short review of direct influences and the basic principles behind the model, before each value is systematically defined.

3.1. Background of the model

17 The present model is based on the notion of discourse domains, which is directly taken from Redeker (1990): she distinguishes between ideational, rhetorical and sequential domains of discourse structure, which can be targeted by DMs depending on the type of elements that are connected. Full definitions will be provided in Section 3.3. González (2004 and 2005) added a fourth component to Redeker's tripartite classification, viz. the interpersonal domain. She also provides a list of functions that each domain includes, such as "conclusion" or "justification". Such an approach in domains and functions is also found in Cuenca (2013), who makes a similar distinction between propositional, modal and structural meanings. According to Cuenca (2013), different types of DMs express different types of meanings (e.g., conjunctions specialize in propositional meanings). These proposals are all corpus-based. However, they are not annotation models *per se*, in that they do not provide operational guidelines on how to systematically apply domains and functions to corpus data.

18 Crible (2017) started from these proposals, and those discussed in the previous sections, and designed an annotation taxonomy for DMs in spoken English and French, which was the basis for our present model. She aimed to address the lack of models specifically designed for discourse *markers* (and not discourse *relations*) in speech, with the additional functions and challenges that they present in this modality, in order to provide a comprehensive quantitative study of DMs with a broad coverage of their types and their functions (as opposed to specific case studies). In its original version, Crible's taxonomy included thirty functions, which are classified across González's (2005) four domains. For instance, the ideational includes the functions of "cause", "condition" and "temporal", while the sequential domain includes "topic-shift", "opening boundary", or "quoting", among others. Each function belongs to one domain only, so that the approach is similar to the PDTB, where a generic class (here, domains) is further refined into a specific function. The distinction between objective and subjective variants of the same relation, which is central in the CCR, only applies to some specific functions with an equivalent in another domain, and therefore a different label: ideational "cause" is the objective equivalent of rhetorical "motivation"; "condition" is the equivalent of "relevance", etc.

19 Crible (2017) reported an intra-rater reliability score of 84% (κ = 0.779) on domains and 75.8% (κ = 0.74) on functions on a sample of 1,194 annotated tokens. However, in annotation experiments involving multiple naive and expert annotators (Crible & Degand, 2019), we found much lower scores of agreement and suggested reducing the number of labels in the taxonomy in order to enhance the replicability

of the annotation. In addition, from a theoretical point of view, in Crible's original proposal, the semantic links between similar functions in different domains are not visible since they receive distinct labels. For instance, when the DM *because* is used with a causal meaning in the ideational domain, it is labeled "cause", while it receives the label "motivation" in the rhetorical domain, thus hiding the shared causal meaning. To address this issue, we introduced in Crible and Degand (2019) the idea of domains and functions as two independent layers of pragmatic information, where any function can combine with any domain. We reported on encouraging agreement results on a small sample of DMs using a preliminary revision of Crible's (2017) taxonomy. The present model in its final form takes up our recommendations of methodological replicability and theoretical adequacy, and is presented in detail in the following sections.

3.2. Principles of the model

20 Compared to Crible's (2017) original taxonomy, the main differences with the present model are that the number of functions is reduced by half from thirty to fifteen and that any function can combine with any domain. For instance, "addition" can be ideational, rhetorical, sequential or interpersonal depending on its contextual interpretation (see examples in Section 4.3 and in Appendix 2). This major change was intended to emphasize the prominent role of domains in the pragmatic variation of DMs: it is not only the specific function or discourse relation expressed by the DM that can change, but it is also the type of connected elements or the speaker's intention that vary. Two utterances can be linked through "addition" either to connect facts (ideational), to serve argumentative purposes (rhetorical), to signal discourse continuity (sequential) or to create complicity with the interlocutor (interpersonal). In other words, the model aims at accounting for the meaning and function of the DMs, as well as for the speaker's communicative intention when using them.

21 As a result, the model now allows for 60 (4 × 15) possible domain-function combinations to be applied to DMs. This is in fact more than the thirty functions defined in Crible's model, since it uncovers more functional variants by refining the objective-subjective distinction (as used in the CCR) and by expanding it to more types of functions. As such, the present model places particular emphasis on the role of discourse domains in DM use. This extension of polyfunctionality to more domains, which themselves apply to more functions, does not necessarily mean that the revised model is less economical than Crible's original one, since the two layers are now independent and the decision process is split in two. The values from each level are therefore defined separately with operational criteria (Appendix 1), in addition to examples for each possible combination (Appendix 2). Section 4.2 reports on inter-annotator agreement scores.

22 We should note that this first and main principle, i.e., any function can combine with any domain, is a theoretical possibility. In our data, a limited number of functions are domain-specific (occurring in only one domain), and some only have variants in two or three domains instead of four. However, this limitation only

applies to the languages and data types that we have analyzed so far (contemporary spoken English and French), and we do not exclude the possibility that more combinations could be attested in other languages or other registers.

23 Another principle of this model is that the annotator can start at either level (domains first or functions first), thus stressing the independence of the two levels of annotation. Furthermore, the annotator assigns one and only one value by level. In other words, we refrain from using double labels (e.g., "consequence-topic"), which are often used in cases of doubt but which are complex to handle quantitatively. We believe that the precise criteria and examples provided in the guidelines should prevent hesitation between values, and systematic biases can be put in place to resolve recurrent ambiguities if necessary [2]. Other cases where double labels are sometimes useful (e.g., a consequence marker with an additional topic-structuring role) can actually often be re-analyzed as the combination of a function and a domain, in our framework (e.g., "sequential consequence"). While this goes against the suggestion in Crible and Cuenca (2017), we believe that double labels are no longer needed in the perspective of efficient annotation and quantitative analysis.

24 Regarding the disambiguation process, in this model we not only pay attention to the basic "dictionary" meaning(s) of the DM (*so* is a marker of "consequence") but also take into account any contextual cue in its interpretation (*so* can express "specification" when it introduces more specific information than in the previous utterance). This means that *and* is not always additive, *but* is not always contrastive, etc., in accordance with the high polyfunctionality of DMs.

25 To sum up, the present model takes up principles from a variety of previous approaches. From the PDTB, we retain the definition of some functions and the operational way in which the annotation guidelines are designed. From the CCR, we reproduce the combinatory approach whereby different layers of information (four in the CCR, two in our model) merge into specifying the particular function. We share with segmentation approaches the attention to the role of context and specifically of units, taking into account hierarchically larger units of speech such as turns and topics. From Crible (2017), we have kept the definition of the domains and some of the functions. All domains and functions will now be defined in the next section.

3.3. Values and definitions

26 We start with the four domains and their definition in Crible (2017: 253), which we further operationalize.

- The *ideational domain* "is linked to states of affairs in the world, semantic relations between external events". In other words, the relation between the

2. We suggest biases in favor of the basic (most frequent) meaning of the DM in case of doubt between two functions. For instance, in a given use of the DM *so*, the annotator might prefer to assign the label "consequence" rather than "topic", since the former is closer to the basic dictionary meaning of the DM.

two discourse objects exists independently in the real world. It corresponds to objective relations and presents the lowest degree of speaker involvement (Pander Maat & Degand, 2001). Operationalization: the arguments of the relation are incompatible with opinionating expressions.

- The *rhetorical domain* "is linked to the speaker's meta-comments on the on-going speech and also includes relations between epistemic or speech-act events". It corresponds to subjective relations and always involves the speaker's attitude or reasoning. Operationalization: the relation needs to be reconstructed with some distance from the content of the segments instead of targeting the contents proper, e.g., referring to the speaker's intentions or beliefs.
- The *sequential domain* "is linked to the structuring of local and global discourse segments such as topics and turns". This means that local management of smaller units (hesitation breaks, other types of filled pauses) will be included in this domain, along with more structural functions such as turn-taking or topic-shifting. Sequential functions explicitly signal the progressing steps of speech and thought. Operationalization: leaving out the DM makes the discourse flow and structure less explicit.
- The *interpersonal domain* "is linked to the interactive management of the exchange and the speaker-hearer relationship". Interpersonal DMs have a phatic function to call for attention or to manifest understanding. Operationalization: the segment cannot be reconstructed without explicitly calling on the addressee.

27 We now turn to the core meaning of each function. This is the invariant meaning aspect, which is then specified in one of the four domains. For most discourse relations, the definition is based on the PDTB guidelines (Prasad et al., 2007).

- *Addition* (ADD): the marker signals that the second segment provides discourse-new information that is related to (but different from) the first.
- *Alternative* (ALT): the marker signals that the segments are alternative situations, exclusive or not. The two units can replace each other.
- *Cause* (CAU): the marker signals that the segment it connects causally explains the situation in the other segment.
- *Concession* (CCS): the marker signals that the segment it connects denies one or several expectations related to the other segment.
- *Condition* (CND): the marker signals that the segment it connects is the condition for the truth or relevance of the other segment.
- *Consequence* (CSQ): the marker signals that the situation in the segment it connects is the result of the situation in the other segment.

- *Contrast* (CTR): the marker signals that there is a shared property between the two segments and that they differ with respect to this property, without any causal inference.
- *Hedging* (HDG): the marker signals some approximation.
- *Monitoring* (MNT): the marker signals the speaker's intent to control the discourse flow.
- *Specification* (SPE): the marker signals that the segment it connects elaborates on the previous segment by giving more detailed information or an example.
- *Temporal* (TMP): the marker signals that the situations in the two segments are chronologically ordered.
- *Agreeing* (AGR): the marker signals agreement with the other speaker.
- *Disagreeing* (DIS): the marker signals disagreement with the other speaker.
- *Topic* (TOP): the marker signals a start of topic, change of topic or return to a previous topic within or between turns. A distant connection to the previous context can remain, with a shift in focus.
- *Quoting* (QUO): the marker introduces (pseudo-)reported speech.

28 The last three functions (disagreeing, topic and quoting) are domain-specific in our data: the first one occurs exclusively in the interpersonal domain, while topic and quoting are always sequential. An overview of all possible combinations between domains and functions can be found in Appendix 1.

4. The proposal in practice

29 The annotation model presented above was established on the basis of a corpus-based study, in which we tested the taxonomy and refined the definitions. In this section, we present the data used in this study, inter-annotator agreement scores calculated on this sample, and distribution results for domains and functions.

4.1. Data used in this study

30 For this study, we took a sample of conversational French from the LOCAS-F corpus (Degand et al., 2014). Specifically, we used three formal and three informal conversations, amounting to 7,545 words in the corpus (about 25 minutes of recordings). The transcripts are sound-aligned and we used the audio in the annotation process, under the Praat software (Boersma & Weenink, 2017). In this data, DMs had already been manually identified, following criteria of syntactic optionality, weak clause association, high degree of grammaticalization, discourse-level scope and procedural meaning (Crible, 2017; Tanguy et al., 2012). A total of 423 DM tokens were annotated. The full list of the 33 DM types is the following: *allez*, *alors*, *après*,

au fond, *ben*, *bien que*, *bon*, *bref*, *donc*, *eh ben*, *en fait*, *en même temps*, *en plus*, *encore que*, *enfin*, *et*, *et puis*, *hein*, *là*, *maintenant*, *mais*, *même que*, *ou*, *ou alors*, *parce que*, *pourtant*, *puis*, *quand même*, *quoi*, *quoique*, *sinon*, *tu vois*, *voilà*.

31 Appendix 3 reproduces the list of all the DMs in the sample with their annotated domains and functions. It should be noted that we considered some complex DMs as one unit (e.g., *et puis*, *eh ben*) when they were fixed, their order of appearance could not be reversed and they performed one joint function (see Cuenca & Crible, 2019, for co-occurrence criteria).

4.2. Inter-annotator agreement

32 All DMs were coded independently by two expert annotators (the authors). The following agreement scores were computed after the first round of annotation, when we were still working on refining the definitions and criteria of the taxonomy. Agreement at the domain level is 71.16% (κ = 0.55), while at the function level it reaches 80.36% (κ = 0.655). Overall, we simultaneously agreed on both the domain and the function on 57.45% of the data.

33 We can first observe that these scores are reversed compared to those reported by Crible and Degand (2019), with a higher agreement on functions than on domains: this confirms that this new model puts the emphasis on the variation brought about by domains, which are therefore more challenging to annotate. A more qualitative analysis of the disagreements revealed that the agreement reaches about 50% for the ideational domain: this is due to a confusion between the ideational and the sequential domains for additive and temporal uses of *et* and *et puis*, as well as a confusion between ideational and rhetorical consequence for *donc*. The other three domains reach agreement in around 75%.

34 For the functions, only two labels present more cases of disagreement than of agreement, namely "specification" – mostly due to three ambiguous DMs, namely *en fait* [in fact] (concession/specification), *enfin* [rather] (alternative/specification) and *donc* [so] (consequence/specification) – and "topic", where uncertainty mostly occurred with the additive function of *et* [and]. Most other functions are quite straightforward, especially "monitoring", "concession", "addition" and "consequence", which all correspond to the core meaning of high-frequency DMs (*hein* [right], *mais* [but], *et* [and], and *donc* [so], respectively).

35 In sum, we observe that the agreement scores (both percentage and kappa) for the functions are much higher than those under Crible's (2017) annotation model. They are similar to those reported for the PDTB 2.0 at "type" level, which is the closest to our taxonomy (84% in Prasad et al., 2008: 2965). Agreement on domains cannot be compared with Crible (2017), since domains and functions are not independent in the latter. While the scores for both levels are lower than Spooren and Degand's (2010) recommended 0.7 kappa threshold, we would like to point that disagreements are mostly due to a small number of problematic expressions.

36 We will now proceed to the results of the distribution of domains and functions of DMs in our sample of conversational French.

4.3. Domain-function combinations in the corpus

37 In the sample, thirty-two different domain-function combinations were found. Four functions have variants in all four domains, namely "addition", "alternative", "concession" and "consequence". Consider the following examples of the latter:

[1] euh Dreyfus donc euh a valu on va dire à Zola euh de s/ d'émigrer en en Angleterre pourquoi parce qu'on on lui a reproché son intervention peut-être un peu trop radicale euh et **donc** Zola euh Zola va devoir partir euh parce que peut-être a-t-il été trop franc

'uh Dreyfus caused let's say Zola uh to emigrate to England why because he was criticized for his intervention maybe a little too radical uh and **donc** [so] Zola uh Zola will have to leave uh because maybe he was too frank'

[2] en gros on en a 256 _ **donc** on a de la marge quoi hein

'in sum we have 256 **donc** [so] we have plenty right'

[3] et donc voilà **donc** euh _ suite à ça ben j'avais con/ j'ai continué les cours et puis euh

'and so there **donc** [so] uh after that well I continued my studies and then uh'

[4] <speaker1> euh enfin je n- ça ne me convenait pas **donc** euh <speaker2> et qu'est-ce qui s'est passé?

'<speaker1> uh well I it didn't work for me **donc** [so] uh <speaker2> and what happened?'

38 In these four examples, *donc* always expresses the relation of "consequence" but each time in a different domain. In [1], the fact that Zola moved to England is the direct, factual consequence of his involvement in the Dreyfus case. In [2], the speaker concludes that "they have plenty" on the basis of a fact ("we have 256"), and this conclusion uses evaluative language as well as other DMs (*quoi* [you know], *hein* [right]), which testifies to its epistemic, rhetorical nature. In [3], *donc* is used in the context of hesitations and helps the speaker restart after a short interruption, taking up her previous narrative. In [4], the consequence is left open, to be reconstructed by the other speaker, as signaled by the turn-final position and the suspensive intonation.

39 From a comparative perspective, these four examples of *donc* would have received completely different treatments in other frameworks. According to the PDTB 3.0, Example [1] would be Result, [2] Result + Belief. With CCR primitives, [1] is causal, positive, basic and objective while [2] is causal, positive, basic and subjective. Both PDTB and CCR would probably not have annotated [3] and [4] at all since they are not strictly connecting segments. In the Val.Es.Co model, [1]-[3] would be considered textual adjacent subacts, while [4] would probably be interpersonal, due to its final position in the intervention/turn. We can see that, with our model,

we can cover all uses of *donc*, including non-connective ones, and make further distinctions that tripartite models for spoken discourse do not propose.

40 In the data, some combinations, such as ideational cause or ideational condition, were not attested, although we have found cases in other corpora. Other functions have only two or three variants, which is in line with our expectations based on our intuitive knowledge of the language and of the particular functions. For instance, "monitoring" has a sequential (Example [5]) and an interpersonal use (Example [6]), and we cannot think of ideational or rhetorical uses of this function – at least not in contemporary French.

[5] c'était pas du tout euh ce qui me convenait et euh _ **ben** euh enfin j'ai arrêté euh l'année passée
'it wasn't right for me at all and uh **ben** [well] uh I mean I stopped uh last year'

[6] j'avais déjà hésité **hein** donc entre euh _ enfin entre institutrice primaire ou euh GRH
'I had already hesitated **hein** [you know] so between uh well between school teacher or HR'

41 As a reminder, the taxonomy also includes three domain-specific functions, that is, functions that only combine with one domain as far as we know ("topic" and "quoting" in the sequential domain and "disagreeing" in the interpersonal domain).

42 The list of attested combinations provided in Appendix 2 includes examples taken from other data, when such cases were not found in the present sample of conversational French. The list contains 42 attested possibilities, including 10 that were not found in our sample but were retrieved from other sources ("rhetorical agreeing", "interpersonal agreeing", "ideational cause", "sequential cause", "ideational condition", "rhetorical contrast", "sequential hedging", "interpersonal hedging", "rhetorical temporal", "interpersonal specification").

4.4. Distribution of domains and function

43 The sequential domain is the most frequent category in the sample, with 48.7% of the occurrences, followed by rhetorical uses (30.3%), as shown in Table 1. The ideational and interpersonal domains have a similar frequency, much lower than the other two (around 10%).

44 This suggests that DMs in French conversations are mostly used to structure discourse (sequential) and to convey the speaker's attitude (rhetorical), rather than to express facts (ideational) or to address the interlocutor directly (interpersonal). The low frequency of the latter category may seem surprising in the highly interactive genre of conversation, but can be explained by the small number of typically interpersonal DMs in French (mainly *hein* [right] and *tu vois* [you see]) and by the peripheral, emerging status of interpersonal variants of discourse relations (as in Example [4] above). These results regarding the distribution of domains are in line with Crible (2017) where the previous version of the taxonomy was applied to a much larger corpus.

45 The top three most frequent functions in our sample are "monitoring" (100 cases, mostly *bon* [well]), "concession" (75, mostly *mais* [but]) and "addition" (72, mostly *et* [and]). Hence, speakers mostly resort to DMs to help manage the discourse flow, to add information and to nuance or contradict. The complete distribution can be found in Table 2.

Domains	**Absolute frequency**	**%**
Sequential	206	48.7
Rhetorical	128	30.3
Ideational	43	10.2
Interpersonal	46	10.9
Total	**423**	**100**

Table 1 – Distribution of domains in the sample

Functions	**Absolute frequency**	**%**
Monitoring	100	23.6
Concession	75	17.7
Addition	72	17.0
Consequence	50	11.8
Specification	34	8.0
Alternative	32	7.6
Temporal	20	4.7
Cause	17	4.0
Topic	13	3.1
Contrast	4	0.95
Condition	2	0.5
Quoting	2	0.5
Hedging	1	0.2
Disagreeing	1	0.2
Total	**423**	**100**

Table 2 – Distribution of functions in the sample

46 There seems to be no association between the frequency of a given relation and the number of domains in which it can be expressed: "monitoring" only has sequential and interpersonal variants and is the most frequent label, whereas "hedging", second-to-last, can be rhetorical, sequential or interpersonal (although the latter two are not attested in the present sample).

47 Taking domains and functions together, the top five most frequent combinations are "sequential monitoring" (Example [7]), "sequential addition" (Example [8]), "rhetorical concession" (Example [9]), "interpersonal monitoring" (Example [10]) and "rhetorical consequence" (Example [11]).

[7] écoute euh _ **ben** euh je sais pas ils faisaient les cons
'look uh **ben** [well] uh I don't know they were acting stupid'

[8] **puis** euh _ **puis** mais euh [laughter] _ mais bon qu'il dit euh _ c'est quand même pas pas [laughter] _ **et** euh _ c'est quand même euh mais ouais elle est quand même trop petite
'**puis** [then] uh **puis** [then] but uh but well he says uh it's not not **et** [and] uh it's still uh but yeah she's too short'

[9] on voulait aller à la séance de 20 h 50 _ **mais** euh c'était bourré massacre
'we wanted to go to the 8:50 show **mais** [but] uh it was fully booked'

[10] pour finir on se dit ben on va aller voir un autre film on n'allait pas _ enfin on va aller voir un autre film tant pis **hein**
'in the end we say well we'll see another movie we were not going to I mean we'll see another movie too bad **hein** [right]'

[11] j'étais là bon _ les gars je vous ramène **alors** soyez calmes
'I was like well guys I'm taking you home **alors** [so] be quiet'

48 How frequent a domain of use is depends on the function: "monitoring", "addition", "specification" and "alternative" are mostly sequential; "concession", "consequence" and "cause" are mostly rhetorical; "temporal" and "contrast" are mostly ideational. It is likely that these preferences will vary across genres (prepared monologue vs. spontaneous conversation), modalities (speech vs. writing) and languages, although further analyses are needed to support this suggestion.

49 Turning to the DMs in the sample, only four types were found to express each of the four domains: *alors* [well/then] (22 cases), *donc* [so] (45 cases), *et* [and] (65 cases) and *mais* [but] (68 cases). These DMs are amongst the most frequent in the sample and all correspond to basic connectives, which express typical discourse relations as well as more interactional functions. However, diversity of domains does not necessarily imply diversity of functions: *ben* [well] (three domains) has been assigned six different function labels, whereas *ou* [or] (three domains also) only

expresses one function (cf. Appendix 3). It remains that *alors*, *et* and *mais* combine a high polyfunctionality in terms of domains (all four are attested) and in terms of functions (six, six and four, respectively), leaving only *donc* with a narrower range of functions ("consequence" and "specification").

5. Discussion: to each their own?

50 Our annotation model, in which domains and functions are independent layers of semantic-pragmatic information, allowed us to describe the distribution of the DM category in a sample of conversational French by providing a fine-grained portrait of their use in spoken language. We not only showed which main aspects of discourse are targeted by the speakers (facts, ideas, structure, exchange) but also through which particular functions they do so (discourse relations such as "cause" or "contrast", speech-specific uses such as "monitoring" or "hedging"). This independent combination also allowed us to identify polyfunctional DM expressions, distinguishing between multi-domain and multi-function types.

51 The present proposal relates to other previous approaches. A version of our four domains was already present in Hovy (1995), although only in relation to simultaneous multifunctionality: the same utterance takes information from the "semantics of the message" (cf. ideational), the "interpersonal speech acts" (cf. interpersonal), "knowledge about stylistic preferability" (cf. rhetorical) and "guidance information" about theme, focus or topic (cf. sequential) (Hovy, 1995: 3). Similarly, Petukhova and Bunt (2009) refer to multiple dimensions only to explain the simultaneous multifunctionality of DMs. Schiffrin (2006) combines monosemy and discourse domains. Overall, the individual components of the present framework are not new, but what is innovative is their combination in a unified and operational model, showing what differs between these components, how they relate to each other and how they apply to DMs.

52 This does not mean, however, that the model we have introduced in this paper is the most efficient or most relevant framework for all research purposes. The point of research is always to overcome previous limitations. Nevertheless, we acknowledge that different approaches to discourse annotation are equally or perhaps more suitable than ours, provided they are methodologically reliable and answer the given research question with the accurate degree of precision. For instance, case studies on particular DMs may require a more fine-grained taxonomy of functions than our list of 15 labels, while some domains and functions will probably never be assigned in studies on written language. Scholars interested in syntax and/or prosody will require some segmentation system, others will want to account for implicit discourse relations, or to focus on pragmatic distinctions which are empirically tested and "cognitively real" (Sanders & Canestrelli, 2012: 211). Cartoni et al. (2013: 81) already noted that "[t]he ideal granularity of the taxonomy is probably not universal but strongly depends on the goal of the annotation", and we fully support this view.

53 In addition, our model presents a number of limitations of its own. Firstly, our inter-annotator agreement results, measured at an early stage of the operationalization of the coding schemes, are rather low, although comparable to other frameworks in the field. The model does not include a procedure for the annotation of implicit relations, even though that would be theoretically possible with the present taxonomy. It does not specify either the order of the segments, which is more systematically included in the PDTB 3.0 and the CCR. Our decision to avoid double labels may be problematic for some cases (e.g., English *as* which is often simultaneously temporal and causal), although this phenomenon is quite restricted. Finally, there is some statistical association between the type of unit and the function of the DM (sequential uses in particular are related to larger discourse units such as turns or topics), which may suggest some degree of conflation between sense disambiguation and segmentation, even if such association is not systematic and does not apply to all functions and domains.

54 Nevertheless, our approach presents a number of specificities and benefits. Chief among them, the combination of domains and functions as independent dimensions is an innovative take on DM polyfunctionality since it extends what previous models have proposed so far. In particular, the model acknowledges that DM functions vary beyond the binary objective-vs.-subjective divide, and this extended variation is applied to more functions than the discourse relations to which it is traditionally restricted. Our framework (or previous working versions of it) has already been applied to different languages (French, English, Polish, Spanish) and modalities (spoken, written, signed) and is therefore well suited for crosslinguistic studies (e.g., Crible et al., 2019; Degand et al., 2018 and in prep.). By starting from the DM itself instead of the relation, it accounts for additional functions of DMs in conversational data, with the rigor and systematicity that are typical of frameworks applied to written data. As such, this taxonomy can fruitfully be combined with other frameworks, for instance by identifying alternative lexicalizations as in the PDTB 3.0, by mapping our functions and domains to the cognitive primitives in the CCR, or by applying systematic segmentation in discourse units *à la* Val.Es.Co. It can also be complemented by more fine-grained, DM-specific analyses of particular expressions.

55 In sum, we would like to call for more research effort striving towards modular discourse models that can apply to many languages and DMs, to both speech and writing, and to many research questions, instead of multiplying marker-specific proposals and thus contributing to the lack of interoperability (or should we say, chaos) in the field of corpus-based discourse analysis. Such a unifying goal (cf. also Sanders et al., 2018) may seem idealistic, but we certainly hope that the present proposal, with its independent dimensions, constitutes a useful addition and complements previous frameworks which share the same goals of interoperability and large coverage of linguistic phenomena, albeit within the range of their own theoretical possibilities.

References

Boersma, P. & Weenink, D. 2017. Praat: Doing Phonetics by Computer. Computer program. Version 6.0.29. URL: http://www.praat.org/.

Briz, A. & Pons, S. 2010. Unidades, marcadores discursivos y posición. In Ó. Loureda Lamas & E. Acín Villa (eds.), *Los estudios sobre marcadores del discurso en español, hoy*. Madrid: Arco Libros: 327-358.

Buysse, L. 2012. *So* as a Multifunctional Discourse Marker in Native and Learner Speech. *Journal of Pragmatics* 44 (13): 1764-1782.

Cartoni, B., Zufferey, S. & Meyer, T. 2013. Annotating the Meaning of Discourse Connectives by Looking at their Translation: the Translation-Spotting Technique. *Dialogue and Discourse* 4 (2): 65-86.

Crible, L. 2017. Discourse Markers and (Dis)fluencies in English and French: Variation and Combination in the DisFrEn Corpus. *International Journal of Corpus Linguistics* 22 (2): 242-269.

Crible, L., Abuczki, A., Burkšaitienė, N., Furkó, P., Nedoluzhko, A., Rackevičienė, S., Valūnaitė Oleškevičienė, G. & Zikánová, Š. 2019. Functions and Translations of Discourse Markers in TED Talks: A Parallel Corpus Study of Underspecification in Five Languages. *Journal of Pragmatics* 142: 139-155.

Crible, L. & Cuenca, M.J. 2017. Discourse Markers in Speech. Characteristics and Challenges for Corpus Annotation. *Dialogue and Discourse* 8 (2): 149-166.

Crible, L. & Degand, L. 2019. Reliability vs. Granularity in Discourse Annotation: What Is the Trade-off? *Corpus Linguistics and Linguistic Theory* 15 (1): 71-99.

Cuenca, M.J. 2013. The Fuzzy Boundaries between Discourse Marking and Modal Marking. In L. Degand, B. Cornillie & P. Pietrandrea (eds.), *Discourse Markers and Modal Particles. Categorization and Description*. Amsterdam – Philadelphia: J. Benjamins: 191-216.

Cuenca, M.J. & Crible, L. 2019. Co-occurrence of Discourse Markers in English: From Juxtaposition to Composition. *Journal of Pragmatics* 140: 171-184.

Degand, L. 2014. "So Very Fast Then" Discourse Markers at Left and Right Periphery in Spoken French. In K. Beeching & U. Detges (eds.), *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*. Leiden – Boston: Brill: 151-178.

Degand, L., Broisson, Z., Crible, L. & Grzech, K. in preparation. Measuring Cross-linguistic Variation in Discourse Markers.

Degand, L., Crible, L. & Grzech, K. 2018. A Multi-dimensional, Multi-functional and Multilingual Account of Discourse Marker Variation. In *DIPVAC4: Discourse-Pragmatic Variation and Change (University of Helsinki, 28-30 May 2018)*. Available online: http://hdl.handle.net/2078.1/200540.

Degand, L., Martin, L.J. & Simon, A.-C. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer & S. Prévost (eds.), *SHS Web of Conferences. Actes du 4e congrès mondial de Linguistique française – CMLF 2014 (Berlin, 19-23 juillet 2014)*. Les Ulis: EDP Sciences. Vol. 8: 2613-2625. Available online: https://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01211.pdf.

Demirşahin, I. & Zeyrek, D. 2014. Annotating Discourse Connectives in Spoken Turkish. In L. Levin & M. Stede (eds.), *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop (August 23-24, 2014, Dublin)*. Stroudsburg – Dublin: Association for Computational Linguistics – Dublin City University: 105-109. Available online: https://www.aclweb.org/anthology/W14-4916.

Evers-Vermeul, J., Hoek, J. & Scholman, M.C.J. 2017. On Temporality in Discourse Annotation: Theoretical and Practical Considerations. *Dialogue and Discourse* 8 (2): 1-20.

Evers-Vermeul, J. & Sanders, T.J.M. 2009. The Emergence of Dutch Connectives: How Cumulative Cognitive Complexity Explains the Order of Acquisition. *Journal of Child Language* 36 (4): 829-854.

González, M. 2004. *Pragmatic Markers in Oral Narrative: The Case of English and Catalan*. Amsterdam – Philadelphia: J. Benjamins.

González, M. 2005. Pragmatic Markers and Discourse Coherence Relations in English and Catalan Oral Narrative. *Discourse Studies* 7 (1): 53-86.

Heim, J.M. 2019. Turn-Peripheral Management of Common Ground: A Study of Swabian *Gell*. *Journal of Pragmatics* 141: 130-146.

Hovy, E.H. 1995. The Multifunctionality of Discourse Markers. In *Proceedings of the Workshop on Discourse Markers (Egmond ann Zee, The Netherlands)*. 1-12. Available online: https://www.isi.edu/natural-language/people/hovy/papers/95dp-egmond.pdf.

Lee, A., Prasad, R., Webber, B.L. & Joshi, A. 2016. Annotating Discourse Relations with the PDTB Annotator. In H. Watanabe (ed.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Stroudsburg: Association for Computational Linguistics: 121-125. Available online: https://www.aclweb.org/anthology/C16-2026.

Mann, W.C. & Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8 (3): 243-281.

Maschler, Y. 2009. *Metalanguage in Interaction. Hebrew Discourse Markers*. Amsterdam – Philadelphia: J. Benjamins.

Pander Maat, H.L.W. & Degand, L. 2001. Scaling Causal Relations and Connectives in Terms of Speaker Involvement. *Cognitive Linguistics* 12 (3): 211-245.

Petukhova, V. & Bunt, H. 2009. Towards a Multidimensional Semantics of Discourse Markers in Spoken Dialogue. In H. Bunt, V. Petukhova & S. Wubben (eds.), *Proceedings of the 8th International Conference on Computational Semantics – IWCS-8 (January 7-9, 2009, Tilburg, The Netherlands)*. Stroudsburg: Association for Computational Linguistics: 157-168. Available online: https://www.aclweb.org/anthology/W09-3715.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B.L. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation – LREC 2008*. Luxembourg: European Language Resources Association: 2961-2968. Available online: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L. & Webber, B.L. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. IRCS technical reports series 203. Philadelphia: University of Pennsylvania ScholarlyCommons. Available online: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1203&context=ircs_reports.

Prasad, R., Webber, B.L. & Lee, A. 2018. Discourse Annotation in the PDTB: The Next Generation. In H. Bunt (ed.), *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (Santa Fe, New Mexico, USA, August 25, 2018)*. Tilburg: Tilburg Center for Cognition and Communication, Tilburg University: 87-97. Available online: https://www.aclweb.org/anthology/W18-4710.pdf.

Redeker, G. 1990. Ideational and Pragmatic Markers of Discourse Structure. *Journal of Pragmatics* 14 (3): 367-381.

Sanders, T.J.M. & Canestrelli, A.R. 2012. The Processing of Pragmatic Information in Discourse. In H.-J. Schmid (ed.), *Cognitive Pragmatics*. Berlin – Boston: De Gruyter: 201-232.

Sanders, T.J.M., Demberg, V., Hoek, J., Scholman, M.C.J., Asr, F.T., Zufferey, S. & Evers-Vermeul, J. 2018. Unifying Dimensions in Coherence Relations: How Various Annotation Frameworks Are Related. *Corpus Linguistics and Linguistic Theory* (Ahead of print): 1-71.

Sanders, T.J.M., Spooren, W.P.M. & Noordman, L.G.M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15 (1): 1-35.

Sanders, T.J.M., Spooren, W.P.M. & Noordman, L.G.M. 1993. Coherence Relations in a Cognitive Theory of Discourse Representation. *Cognitive Linguistics* 4 (2): 93-133.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schiffrin, D. 2006. Discourse Marker Research and Theory: Revisiting *And*. In K. Fischer (ed.), *Approaches to Discourse Particles*. Amsterdam – London – Paris: Elsevier: 315-338.

Scholman, M.C.J., Evers-Vermeul, J. & Sanders, T.J.M. 2016. A Step-Wise Approach to Discourse Annotation: Towards a Reliable Categorization of Coherence Relations. *Dialogue and Discourse* 7 (2): 1-28.

Spooren, W.P.M. & Degand, L. 2010. Coding Coherence Relations: Reliability and Validity. *Corpus Linguistics and Linguistic Theory* 6 (2): 241-266.

Sprott, R.A. 1992. Children's Use of Discourse Markers in Disputes: Form-Function Relations and Discourse in Child Language. *Discourse Processes* 15 (4): 423-439.

Tanguy, N., Van Damme, T., Degand, L. & Simon, A.-C. 2012. Projet FRFC "Périphérie gauche des unités de discours" – Protocole de codage syntaxique. 1-17. Available online: http://halshs.archives-ouvertes.fr/halshs-00762866.

Tonelli, S., Riccardi, G., Prasad, R. & Joshi, A. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogs. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation – LREC 2010*. Luxembourg: European Language Resources Association: 2084-2090. Available online: http://www.lrec-conf.org/proceedings/lrec2010/pdf/184_Paper.pdf.

Val.Es.Co Group 2014. Las unidades del discurso oral. *Estudios de Lingüistica del Español* 35: 13-73.

Zhang, F., Litman, D. & Forbes Riley, K. 2016. Inferring Discourse Relations from PDTB-Style Discourse Labels for Argumentative Revision Classification. In Y. Matsumoto & R. Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Stroudsburg: Association for Computational Linguistics: 2615-2624. Available online: https://www.aclweb.org/anthology/C16-1246.

Appendices

Appendix 1: annotation scheme

	Ideational (IDE)	**Rhetorical (RHE)**	**Sequential (SEQ)**	**Interpersonal (INT)**
Addition (ADD)	Addition between two facts, usually in single clauses.	Argumentative addition or emphatic effect, typically expressed by "moreover" [*et surtout*] (co-occurrence test: "and moreover", "and on top of that").	Continuity, mere linkage of utterances: the discourse continues with no added meaning, typically in a narrative and/or between larger units (complex idea units, turns).	Addition that echoes/repeats another speaker's words.
Alternative (ALT)	Two competing facts, exclusive alternative "either... or". Includes chosen alternative (*instead*).	Reformulation of two full units, one is preferred by the speaker (the reparans). Paraphrastic or non-paraphrastic. The 2nd introduces a change in meaning (not just a difference in phrasing).	Repair due to a change in phrasing or with incomplete units in a disfluent sequence (no subjective preference because the reparandum is not verbally expressed, the marker just restarts the flow after the interruption).	Other-repair, the reparandum is produced by the other speaker.
Cause (CAU)	The segment introduced by the DM is the logical cause of the other segment, effect-reason relation between facts.	Epistemic or speech-act cause, need to reconstruct a reasoning "I can say this because...".	Cause that also serves a discourse-structuring purpose, such as topic-shift.	Cause that answers a question asked by the other speaker or that responds to the other speaker in any way (for instance with agreeing or disagreeing tone).

Concession (CCS)	Logical counter-expectation between two facts with very limited subjective reasoning.	The concessive link needs to be reconstructed, explicitly involves personal opinions, speech-acts or epistemic assumptions.	There is some opposition between the two arguments but it also performs some structuring function, applies to larger segments, the marker corresponds to a major boundary.	Opposition of opinions, exclusively in a dialogic context.
Condition (CND)	The segment introduced by the DM is the logical condition for the other segment (which is the consequence), includes all subtypes (present, past, etc.) and negative hypothesis ("unless" [*sinon*]). Mainly expressed by conditionals "if", "provided".	The two arguments are not causally related but the segment introduced is what makes the speech-act or epistemic conclusion relevant to the particular context "I can say this only in the context of...".	/	/
Consequence (CSQ)	The segment introduced by the DM is the logical effect or result brought by the first segment (forward causality). Includes purpose relation ("so that"). The inference is very limited to objective facts.	Epistemic or speech-act consequence, including summary with conclusive value, usually taking scope over a large previous context. Strong speaker's appreciation of the causal link between the two segments "I can now say/ conclude that...".	Epistemic or speech-act consequence which also performs some structuring function such as topic-resuming. Major boundary, higher in the discourse hierarchy.	No linguistically expressed consequence, to be reconstructed by the addressee, signals that the interlocutor can take the turn (turn-yielding). Final position not a sufficient criterion.

Contrast (CTR)	Clear opposition between two facts, usually marked by syntactic or lexical devices in addition to the marker (e.g., antonyms). An entity and a property are compared. The property is verbally expressed.	The contrast serves an argumentative purpose, one of the two opposed units is subjectively preferred or more important. Includes corrective uses ("not... but").	Two major segments (e.g., scenarios) are contrasted with a structuring function.	/
Hedging (HDG)	/	Approximation to avoid a literal understanding or because of epistemic uncertainty, refers to the speaker's knowledge.	Approximative marker used to stall, to fill a gap.	Approximation because of politeness or face-threatening material.
Monitoring (MNT)	/	/	Keep control over the turn/ discourse, self-monitoring (former "Punctuation"), usually in contexts of hesitation, stagnation.	Keep control over the interaction, maintain contact with the interlocutor, other-monitoring.
Specification (SPE)	The segment introduced gives more detailed information about the previous segment: a detail or an example. It can be directly subsumed under the previous segment (informational dependence), corresponds to a colon ":".	Addition of a detail which is subjectively appreciated by the speaker (in focus, more important): specification with some stylistic (emphatic) effect.	Addition of a detail or comment which is presented as a parenthetical aside, withdrawn from the linear structure of the discourse. Or specification of a previously introduced referent that opens a new boundary. Or answer to a question.	Addition of a detail or comment as an answer to a question which also conveys some face-saving function.

Temporal (TMP)	The two facts are chronologically related, includes simultaneous, precedence and succession. Bias for temporal in case of conflict with consequence relations ("then").	The two arguments or segments are steps in the argumentation of the discourse, with a cline such that what comes later is stronger. Or speech-act temporal relation.	The two arguments or segments are steps in the chronology of the discourse, similar to bullet points.	/
Agreeing (AGR)	/	Expression of conforming opinion with oneself (no dialogic exchange).	/	Expression of conforming opinion with the addressee.
Disagreeing (DIS)	/	/	/	Expression of discording opinion, when none of the other functions apply.
Topic (TOP)	/	/	Mere marking of topic-shift and topic-resuming, when none of the other functions apply.	/
Quoting (QUO)	/	/	Introducing (pseudo-) reported speech.	Bias for SEQ, but it is somewhat in-between SEQ and INT.

Appendix 2: examples for each attested combination and their English translation

Addition

Ideational:

le grand frère avait un rôle de papa ***et*** *en plus d'être papa il avait un rôle de d'essayer les choses avant nous*

'the big brother had the role of a daddy and in addition to being a daddy he had the role of trying things before us'

Rhetorical:

non je marchais pas ah non non j'ai pas couru (0.180) ***et*** *j'ai fait encore un détour*

'no I wasn't walking ah no no I didn't run (0.180) and I did a detour'

Sequential:

Pacs avait fait une intendance aux baladins (0.780) ***et*** *euh Camille lui dit euh tu oublieras pas de payer*

'Pacs had been working as a steward with the boy scouts (0.780) and uh Camille told him uh you won't forget to pay'

Interpersonal:

<spk1> tu dis euh cheese pour le cliché et genre euh un peu pour se cacher <spk2> ***et*** *un peu pour se cacher aussi ouai*

'<spk1> you say uh cheese for the cliché and like uh a little to hide yourself <spk2> and a little to hide myself too yeah'

Alternative

Ideational:

on est plusieurs ***ou*** *tu me vouvoies?*

'there are several of us or you're being polite?'

Rhetorical:

c'est pas pour ça qu'on fait de la musique mais c'est ***enfin*** *c'est pas pour être reconnu dans la rue*

'that's not why we're in music but it's I mean it's not to be recognized in the street'

Sequential:

euh ben j'ai fait euh deux ans ***enfin*** *ma première et ma deuxième euh d'institutrice euh primaire*

'well I studied uh two years well my first and my second uh of primary school teacher'

Interpersonal:

<spk1> j'avais repris euh des études en gestion des ressources humaines […] <spk2> directement après? <spk1> ben euh ***enfin*** *j'ai arrêté euh l'année passée euh avril et euh […] l'année scolaire suivante*

'<spk1> I was back in school studying human resources management [...] <spk2> right after? <spk1> well uh actually I stopped uh last year uh in April and uh [...] the next year'

Cause

Ideational:

*les monos voulaient pas rester **parce qu'**elles avaient trop peur*
'the instructors didn't want to stay because they were too scared'

Rhetorical:

*c'est le titre d'un d'un assez long poème **puisqu'**il fait cinquante pages*
'it's the title of a rather long poem since it is fifty pages long'

Sequential:

*<spk1> c'était pas mieux c'était totalement différent <spk2> **parce que** vous quand vous avez 15 ans vous vivez la guerre?*
'<spk1> it wasn't better it was totally different <spk2> because you when you are 15 you are living the war?'

Interpersonal:

*ouais hein c'est vrai **parce que** objectivement tout homme malheureux que vous êtes vous jouissez quand même d'un grand succès*
'Yeah right it's true because objectively however unhappy you are you still have a huge success'

Concession

Ideational:

*elle devait partir le lendemain **mais** elle n'est jamais partie*
'she was supposed to leave the next day but she never left'

Rhetorical:

***si** la démocratie est un mot ancien, ici et maintenant la démocratie signifie la prospérité pour tous*
'while democracy is an old word, here and now democracy means prosperity for all'

Sequential:

*c'était assez comique de les entendre parler comme ça euh des filles (0.690) **mais** euh ouais puis après euh voilà quoi*
'it was quite funny hearing them talk like that uh about the girls (0.690) but um yeah then after uh that's it'

Interpersonal:

*cet auditeur euh vigilant il va vous dire tiens euh encore Jean d'Ormesson **mais** on entend Jean d'Ormesson à chaque automne*
'this careful listener he's going to tell you uh Jean d'Ormesson again but we hear Jean d'Ormesson every autumn'

Condition

Ideational:

***si** nous avons la responsabilité du pays nous donnerons des papiers à tous ceux qui n'en ont pas*

'if we have the responsibility of the country we will give papers to all who don't have any'

Rhetorical:

*il devait y avoir une porte alors **si** c'est la sacristie*
'there must have been a door then if it was the sacristy'

Consequence

Ideational:

*l'identité est un effet structurel un rapport et **du coup** elle mobilise obligatoirement des signes visibles*
'identity is a structural effect a relationship and as a result it necessarily involves visible signs'

Rhetorical:

*en gros on en a 256 (0.990) **donc** on a de la marge quoi hein*
'we have about 256 of them (0.990) so we have plenty right'

Sequential:

*et donc voilà **donc** euh _ suite à ça ben j'avais con/ j'ai continué les cours et puis euh*
'and so there so uh after that well I continued my studies and then uh'

Interpersonal:

*<spk1> euh _ enfin je n- ça ne me convenait pas **donc** euh <spk2> et qu'est-ce qui s'est passé?*
'<spk1> uh well it wasn't right for me so uh <spk2> and what happened?'

Contrast

Ideational:

*Johnny Halliday ils connaissent pas **mais** moi ils connaissent hein*
'they didn't know Johnny Halliday but they knew me right'

Rhetorical:

*on ne le conçoit pas qu'un éloge soit écrit euh de façon neutre pâle fade impersonnelle et et et **au contraire** l'éloge demande qu'on soit engagé complètement*
'we can't imagine that a eulogy be written in a neutral pale impersonal way and and and on the contrary the eulogy requires one to be completely committed'

Sequential:

*il y en a un qui s'est branché sur les mondains et les histoires de tatas et de pédés **et puis** il y en a un autre qui s'est branché sur des histoires de misère*
'one was interested in the social elite and stories of faggots and then the other was interested in stories of misery'

Hedging

Rhetorical:

*après tu as un espèce de bêtisier **là***
'then you have a blooper sort of'

Sequential:

le XIX^e^ siècle est un siècle beaucoup complexe extrêmement complexe par rapport ***je vais dire*** *euh à un XVII^e^ ou un XVIII^e^*

'the 19th century is a century much more complex extremely complex compared to like uh to the 17th or 18th'

Interpersonal:

ils ont été éduqués dans ***je dirais*** *dans le français (0.287) comme vous dites modèle*

'they were educated in I'd say in as you say standard French'

Monitoring

Sequential:

et donc voilà donc euh suite à ça ***ben*** *j'avais con/ j'ai continué les cours*

'and so there so uh after that well I continued my studies'

Interpersonal:

ce n'est pas un mémoire de romane ***hein***

'it's not a MA thesis in Romance philology you know'

Specification

Ideational:

les nobles vont s'engager dans la lit/ dans la lutte politique et sociale par leurs œuvres littéraires et leurs actions politiques bien sûr ***par exemple*** *Hugo et Lamartine sont députés bon ils remplissent des fonctions politiques*

'the nobles will get involved in political and social fights through their literary works and their political actions of course for example Hugo and Lamartine are members of parliament well they have political duties'

Rhetorical:

dès qu'on a un événement de communication on a un style de parole ***et*** *c'est ce style de parole qu'on a essayé de décerner*

'as soon as we have a communicative event we have a speech style and it is this speech style that we tried to define'

Sequential:

tu ne peux avoir qu'une seule (0.770) ***et*** *c'est assez logique (0.510) qu'une seule machine*

'you can only have one (0.770) and it's quite logical (0.510) only one machine'

Interpersonal:

<spk1> j'ai pas le profil <spk2> c'est-à-dire? <spk1> ***bah*** *[laughs] ça c'est bon on ne le dit pas mais je le ressens*

'<spk1> I don't have the profile <spk2> what do you mean? <spk1> well [laughs] that's well they don't say it but I feel it'

Temporal

Ideational:

j'ai continué les cours ***et puis*** *euh arrivée au deuxième stage je n'ai pas euh je n'ai pas entrepris de de le faire*

'I continued my studies and then uh at the second internship I didn't uh I didn't carry it out'

Rhetorical:

dans Voyage au bout de la nuit *il commence une ligne en disant Proust (0.350) mi revenant lui-même (0.290)* ***déjà*** *c'est sublime mi revenant lui-même*
'in *Voyage au bout de la nuit* he starts a line saying Proust (0.350) half ghost himself (0.290) first that's beautiful half ghost himself'

Sequential:

d'abord on commence par _ euh le point de vue politique le point de vue industriel scientifique comme je vous l'ai dit ***et puis*** *on passera au domaine économique*
'first we start with uh the political standpoint the industrial scientific standpoint as I told you and then we will move on to the economic domain'

Agreeing

Rhetorical:

bah dans l'esprit actuel des gens ce genre de bouquin ferait un best-seller et serait accompagné d'une pub gratuite. Pub mauvaise ***certes*** *mais pub quand même*
'well in people's current mood this kind of book would make a best-seller and would be accompanied with free advertising. Bad advertising granted but still advertising'

Interpersonal:

cette demande […] pourrait ***certes*** *permettre de rendre un peu de cohérence à un système de financement actuellement absurde*
'this request could indeed allow us to give back some coherency to a financial system currently absurd'

Disagreeing (Interpersonal)

il dit maintenant on va de nouveau retomber dans un krach // ***ben*** *ça n'a pas autant remonté que ça*
'<spk1> he says now we will fall back again into a crash <spk2> well it didn't recover that much'

Topic (Sequential)

Mme Ebadi a annoncé son intention de contester cette décision (0.510) ***et puis*** *des nouvelles de la santé de Fidel Castro*
'Mrs Ebadi announced her intention to protest this decision (0.510) and now some news of Fidel Castro's health'

Quoting (Sequential)

alors pour finir on s'est dit ***ben*** *on va aller voir un autre film*
'so in the end we thought well we will go see another movie'

Appendix 3: DMs in the sample with their annotated domains and functions

Discourse marker	Domains	Functions
allez (1)	SEQ (1)	QUO (1)
alors (22)	IDE (5), RHE (4), SEQ (12), INT (1)	CSQ (8), ADD (6), TMP (3), ALT (2), SPE (2), TOP (1)
après (1)	SEQ (1)	CTR (1)
au fond (2)	RHE (2)	SPE (1), CCS (1)
ben (31)	RHE (1), SEQ (27), INT (3)	MNT (232), SPE (4), DIS (1), TOP (1), CCS (1), QUO (1)
bien que (1)	RHE (1)	CCS (1)
bon (37)	RHE (1), SEQ (34), INT (2)	MNT (34), ALT (1), TOP (1), SPE (1)
bref (1)	SEQ (1)	MNT (1)
donc (45)	IDE (8), RHE (23), SEQ (13), INT (1)	CSQ (36), SPE (8), ALT (1)
eh ben (4)	SEQ (4)	MNT (3), CSQ (1)
en fait (11)	RHE (9), SEQ (2)	SPE (6), CCS (4), TOP (1)
en même temps (1)	RHE (1)	CCS (1)
en plus (1)	RHE (1)	ADD (1)
encore que (1)	RHE (1)	CCS (1)
enfin (32)	RHE (13), SEQ (18), INT (1)	ALT (22), SPE (5), CSQ (3), MNT (2)
et (65)	IDE (8), RHE (11), SEQ (45), INT (1)	ADD (52), SPE (5), CSQ (3), MNT (2)
et puis (15)	IDE (8), RHE (1), SEQ (6)	TMP (10), ADD (5)
hein (17)	INT (17)	MNT (17)
là (1)	RHE (1)	HDG (1)
maintenant (2)	RHE (1), SEQ (1)	TMP (1), CCS (1)
mais (68)	IDE (6), RHE (31), SEQ (25), INT (6)	CCS (58), SPE (4), CTR (3), ADD (2), TOP (1)
même que (1)	RHE (1)	ADD (1)
ou (5)	IDE (2), RHE (1), SEQ (2)	ALT (5)

ou alors (1)	IDE (1)	ALT (1)
parce que (17)	RHE (16), INT (1)	CAU (17)
pourtant (2)	RHE (2)	CCS (2)
puis (10)	IDE (5), SEQ (5)	TMP (5), ADD (5)
quand même (3)	RHE (3)	CCS (3)
quoi (17)	SEQ (5), INT (12)	MNT (17)
quoique (1)	RHE (1)	CCS (1)
sinon (3)	RHE (2), SEQ (1)	CND (2), TOP (1)
tu vois (1)	INT (1)	MNT (1)
voilà (3)	SEQ (3)	MNT (2), TOP (1)